Comparabilty of GATB scores for immigrants and majority group members: Some Dutch findings

te Nijenhuis, J.; van der Flier, H.

published in
Journal of Applied Psychology
1997

DOI (link to publisher)
10.1037/0021-9010.82.5.675

document version
Publisher's PDF, also known as Version of record

Journal of Applied Psychology
1997, Vol. 82, No. 5, 675-687

0021-9010/97/$3.00

Comparability of GATB Scores for Immigrants and Majority Group Members: Some Dutch Findings

Jan te Nijenhuis and Henk van der Flier
Vrije Universiteit and Dutch Railways

The central question addressed in this article is whether the test scores of immigrants and majority group members reflect the same dimensions. Use was made of scores on the Dutch version of the General Aptitude Test Battery on first-generation immigrants (N = 1,322) and majority group members (N = 806) who applied for blue-collar jobs in the Netherlands. The group differences with respect to the construct validity were small. Spearman's hypothesis that general intelligence is the predominant factor determining the size of the differences between 2 groups was borne out significantly. The test can be put to good use for comparisons within culturally homogeneous groups of non-native-born, non-native-language minorities. Use of the test for comparisons between immigrant and majority group members, however, requires supplementary research.

Following Binet's (Binet & Simon, 1905) example, standardized ability tests were developed in Western countries, and they are now being used all over the world. An important question is whether this use is justifiable on the grounds of the results of empirical research. Because important selection decisions are often made on the basis of test scores, small group differences in validity can have large consequences for groups. Test users are therefore obliged to base their professional judgments on instruments having an established validity. Research on White and non-White populations in North America and Europe shows that standardized ability tests have predictive validity in work and learning situations and that they have construct validity. On the basis of a review of research outcomes in the United States, Jensen (1980) concluded that

> The currently most widely used standardized tests of mental ability—IQ, scholastic aptitude, and achievement tests—are, by and large, *not* biased against any of the native-born English-speaking minority groups on which the amount of research evidence is sufficient for an objective determination of bias, if the tests were in fact biased. (p. ix)

Hunter, Schmidt, and Hunter's (1979) meta-analysis came to similar conclusions. More recently, researchers have been discussing how best to deal with the established mean differences in intelligence between the groups distinguished by Jensen (Arvey & Faley, 1988; Gottfredson, 1994; Hartigan & Wigdor, 1989; Herrnstein & Murray, 1994).

Research conducted with the General Aptitude Test Battery (GATB; U.S. Department of Labor, 1970) has played an important role in recent discussions. Hunter and Hunter (1984) showed that the GATB has high predictive validity. Large-scale research with the GATB has shown no significant differences between the regression lines of the criterion on the predictor for the majority and minority groups (Hunter, 1983b). Finally, a book by Hartigan and Wigdor (1989) on how to deal with group differences in intelligence was based on research with the GATB.

Jensen's (1980) research concerned differences in test scores between native born English-speaking minorities and Whites, which means that his conclusions cannot be extrapolated either to populations in developing countries or to immigrant populations in Western countries.

Cross-cultural psychologists have paid particular attention to the dimensional comparability of measuring instruments. In a review of the literature on standardized ability tests, Vandenberg and Hakstian (1978) discussed the mean values of the congruence coefficient between the factor solutions of groups from Western countries and groups from developing countries. The congruence coef-

Jan te Nijenhuis and Henk van der Flier, Department of Work and Organizational Psychology, Vrije Universiteit, Amsterdam, the Netherlands, and Dutch Railways, Utrecht, the Netherlands.

The Training and Recruitment Foundation of Dutch Railways provided financial support for this research.

An earlier version of this article was presented at the 8th Congress of the Dutch Institute of Psychologists, Nijmegen, the Netherlands, October 1994, and at the 26th International Congress of Psychology, Montréal, Canada, August 1996.

Correspondence concerning this article should be addressed to Jan te Nijenhuis, Department of Work and Organizational Psychology, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, the Netherlands. Electronic mail may be sent via Internet to j.te.nijenhuis@psy.vu.nl.

ficient (Burt, 1948; Tucker, 1951) is a measure of the proportionality of columns, which, applied to factor loadings, is a good measure of equality of interpretation of factors; a value greater than .85 is generally considered to be high. The values of the congruence coefficient were .83 for Numerical Intelligence, .84 for Perceptual Speed, .89 for Verbal Intelligence, and .92 for Visualization. A value of .83 might mean that most of the tests have corresponding loadings and that some of the tests show substantial discrepancies in the factor loadings. A congruence coefficient with a value of .92 might mean that virtually all of the tests have corresponding loadings and that a single subtest shows substantially changed factor loadings. A substantial discrepancy of the factor loadings of a subtest signifies that this subtest does not measure the same dimensions and that it therefore cannot be used for comparisons between groups without encountering problems.

In the last few decades, a growing number of immigrants have become part of the Dutch population. The immigrants in the Netherlands come mainly from Suriname, the Dutch Antilles, Morocco, and Turkey. Suriname is an ex-colony of the Netherlands, and the Dutch Antilles still form a part of the Netherlands, so the majority of the Surinamese and the Dutch Antillians have a good command of the Dutch language. Most of these immigrants work in jobs at the lower end of the labor market, and a relatively large percentage are unemployed. Their mean test results are lower than the mean test results of the Dutch. The use of tests to assess immigrants is being criticized in both the popular and scientific press in the Netherlands. These critics assume that tests are of limited use for assessing persons with a limited knowledge of the Dutch language and culture. However, it is still too early to draw conclusions because in the Netherlands little research has addressed the degree to which test scores for immigrant groups can be compared with those for the majority group.

Our research bears on the current debate on the assessment of non-native-born, non-native-speaking minorities in Europe. It is also relevant for the assessment of immigrants in the United States because of its attention to the influence of language and culture on test scores.

Research Question

Messick (1989) provided a comprehensive definition of validity: "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Research into comparability of test scores between immigrants and majority group members addresses the question of whether the same conclusions can be drawn from the same test scores for immigrants and majority group members or, in other words, whether the validity of the test is the same in both groups. Assessment of the comparability of scores on intelligence tests focuses on two different questions: The first is whether immigrants and majority group members with the same test scores have the same level of intelligence. The second question is whether immigrants and majority group members with the same test scores have the same chance of showing specific criterion behavior in the future. These two questions pertain to the traditional classification between construct validity and predictive validity. Construct validity concerns whether the same dimensions are measured and whether measurement is carried out using the same measurement units at the same level. To assess dimensional comparability, one must check whether tests are connected to relevant constructs in a comparable fashion in the different groups. For this purpose, outcomes of correlational analyses (e.g., correlation and factor matrices) are often compared. The presence or absence of differential item functioning also has implications for construct validity.

The central question addressed in this article is whether the test scores of immigrants and majority group members reflect the same dimensions.

Method

Research Participants

In this project, we used test data on first-generation immigrants and majority group members who applied for blue-collar jobs at the Dutch Railways and regional bus companies in the Netherlands from 1988 until 1992. The jobs were varied and ranged from train cleaner to rail maintenance expert. The application process included a psychological examination, which took place at the Work Conditions Service Unit of the Dutch Railways in 10 centers throughout the Netherlands. A subsample was selected from the majority group in such a way that the distribution with respect to the jobs and regions in this subsample was as close as possible to that in the immigrant group. Table 1 shows the distributions of the groups in terms of demographic

Table 1
Distribution of the Immigrant and Majority Group Members With Respect to Native Country, Size of Subgroup, Percentage of Men, and Age

Native country	*n*	% of men	Mean age
Suriname	535	82.5	30.2
Dutch Antilles	126	83.5	31.2
North Africa	167	97.4	27.4
Turkey	275	96.5	24.0
Other	219	85.5	30.6
Netherlands	806	87.6	28.4

Note. The group of North Africans consists of persons from Morocco, Algeria, Tunisia, Libya, and Egypt.

variables. The immigrant group from countries classified as *other* consists of persons from Asia (with the exception of Turkey, Israel, Japan, and the countries of the former Soviet Union); Africa (with the exception of Suriname); and Central America (with the exception of the Dutch Antilles). In view of the large heterogeneity of this group, its test scores are only reported in analyses of all of the immigrants, treated as a single group. The mean number of years that the immigrants had been residing in the Netherlands at the time of their application was 11.2 years (SD = 6.9 years). This might be regarded as a reasonable amount of time for the majority of the immigrants to absorb the basics of the Dutch language and to become at least basically familiar with Dutch culture. Surinamese and Antillian immigrants are diverse with regard to their ethnicity. The population of Suriname consists mainly of Creoles (persons of African or mixed African and European background) and Asian Indians, with smaller groups of Indonesians, Chinese, Whites, and American Indians. The population of the Dutch Antilles consists mainly of Creoles. Official registration of immigrants with regard to their ethnicity did not take place in the Netherlands at the time the data were collected.

Test

The GATB 1002 B (General Aptitude Test Battery) is a test of general intelligence. All of its eight tests are speeded, and all but one are in multiple-choice format. To describe the GATB subtests, research into hierarchical factor models can be used. Among these models, the Cattell-Gustafsson model (Carroll, 1993; Cattell, 1987; Gustafsson, 1984, 1988) is widely accepted. At the highest level of the hierarchy (stratum III) is general intelligence or g; one level lower (stratum II) are the broad abilities, Fluid Intelligence, Crystallized Intelligence, General Memory and Learning, Broad Visual Perception, Broad Auditory Perception, Broad Retrieval Ability, and Broad Cognitive Speediness or General Psychomotor Speed; one level lower (stratum I) are the narrow abilities, such as Sequential Reasoning, Quantitative Reasoning, Verbal Abilities, Memory Span, Visualization, and Perceptual Speed; at the lowest level of the hierarchy are the specific tests and subtests. Carroll's (1993) terminology is used throughout this article because it is based on the most up-to-date, comprehensive, and influential taxonomy available.

What follows is our classification of the different GATB subtests in terms of Carroll's taxonomy. At stratum I, several tests are unidimensional, whereas others are multidimensional. Three-Dimensional Space measures Visualization, Vocabulary measures Induction and Lexical Knowledge, Arithmetic Reasoning measures Quantitative Reasoning, Computation measures Numerical Ability, Tool Matching measures Perceptual Speed, Form Matching measures Spatial Relations, Name Comparison measures Perceptual Speed and Numerical Ability, and Mark Making measures Aiming. At stratum II, the subtests Three-Dimensional Space, Tool Matching, and Form Matching measure Broad Visual Perception. The subtests Vocabulary and Arithmetic Reason measure both Fluid and Crystallized Abilities. The subtest Computation measures Crystallized Abilities. The subtest Name Comparison measures both Broad Visual Perception and Crystallized Abilities. The subtest Mark Making measures General Psychomotor Speed.

Carroll (1993, p. 597, Table 15.5) reported the central tendencies of loadings of first-order factors on third-order g factors. Carroll (1993, p. 625) stated that the factor General Psychomotor Speed has minimal cognitive content. Considering this information, at stratum III, all of the tests are expected to have low (Mark Making) to high loadings (Vocabulary and Arithmetic Reason) on a factor of general intelligence.

The Dutch version of the test was administered. The subtests Computation, Three-Dimensional Space, Tool Matching, Form Matching, and Mark Making of the Dutch version are literal translations of the American version; the original stimuli were retained. The other subtests contain minor adaptations to the Dutch language and culture such as the introduction of Dutch and other European names in Name Comparison and Dutch money in Arithmetic Reason. As in the original GATB, the Dutch GATB contains three psychomotor subtests; of these three subtests, only Mark Making is used at the Dutch Railways.

A review of test research in the Netherlands by Evers, van Vliet-Mulder, and Ter Laak (1992) showed that the Dutch version of the test has good predictive validity, content validity, and construct validity. Scores on the subtests of the majority group members are therefore generally considered to provide a good indication of their capacities.

Jensen and Weng (1994) stated that the goodness of the g extracted from a set of tests depends on, among other things, (a) the number of tests, (b) the number of different mental abilities represented by the various tests, and (c) the degree to which the different types of tests are equally represented in the set. The g factor varies across different sets of tests to the extent that the sets depart from these criteria. Jensen (1985) showed that the g factor of the GATB is highly, but not perfectly, similar to the General Ability factor reflected in the total score of tests, which are seen as representing g well on the basis of their broad and well-balanced sampling of abilities. The nonperfect correlation can be explained by the fact that the GATB does not measure all broad abilities: It measures Fluid and Crystallized Intelligence, Broad Visual Perception, and General Psychomotor Speed but does not contain tests for General Memory and Learning and Broad Retrieval Ability. Notwithstanding this limited sampling of broad abilities, the GATB has high predictive validity (Hunter & Hunter, 1984).

Statistical Analyses

Means and reliabilities. The deviation of the mean scores of the immigrants on the GATB subtests from the mean scores of the majority group members was calculated in terms of the standard deviation of the majority group. To estimate the reliability of four GATB subtests, Cronbach's alpha was chosen. This method for determining the consistency of items is suitable for power tests. A consequence of testing under time limits is that a large number of the participants will not answer the items at the end of the test. Because the scores of these persons on the last items will all be zero, the correlations between these last items will be high, which will make the alphas spuriously high. The items of the GATB subtests that were answered by close to 90% of the participants of the groups in question were consid-

ered to constitute a power test, on which the alpha was computed.

Dimensional comparability. The dimensional comparability of the subtests for the majority group and the immigrant groups was examined by means of structural equation modeling, using EQS (Bentler, 1989). Several models with increasing degrees of constraint were fitted to the data. The following tests were examined: (a) tests of comparability of covariance matrices, (b) tests of the same number of factors in two groups, and (c) tests of the equality of factor loadings in two groups. The factor model tested across groups was based on Carroll's (1993) version of the Cattell-Gustafsson model and Hunter's (1983a) study of the dimensionality of the GATB and represented a three-factor solution. The first factor is a hybrid of Fluid and Crystallized Intelligence and is called G_H. The second factor is related to Broad Visual Perception and is called G_V. The third factor is related to General Psychomotor Speed and is called G_P.

When working with large samples, even small differences between groups can lead to large chi-square values; these chi-square values will make the small differences significant (Jöreskog, 1969). For that reason, various researchers have suggested additional goodness-of-fit measures, such as the comparative fit index (CFI) (Bentler, 1989), which has been shown to be less susceptible to the effects of sample size than other measures.

Congruence coefficients (Burt, 1948; Tucker, 1951) of factor loadings for pairs of groups were examined to further explore the relationships of factor loadings between groups; a value greater than .85 is generally considered to be high.

Interpretation of a difference in factor loadings as large or small depends on the nature and context of the research. One way of interpreting the size of discrepancies in factor loadings in this research is to compare them with the discrepancies found in research on the dimensional comparability for men and women. In a large-scale study, discrepancies in factor loadings were, on the whole, smaller than .05, with the largest female–male difference being .12 (Carretta & Ree, 1995). A discrepancy in factor loadings for a subtest of more than .10 is therefore called *substantial* in this study.

Differential item functioning. Research into differential item functioning starts with a definition of what constitutes biased and unbiased items. On the basis of statistical procedures that are operationalizations of the definition of item bias, the question of which items are biased can then be addressed. The term *statistical item bias* is used because the bias in question pertains only to statistical deviance; the biased items deviate only in a statistical sense from the other items. Little can be said at this stage about the cause of this statistical deviance. Finally, on the basis of statistical results and other information, hypotheses are formulated about qualities of the tested persons or the items (or both) that might be responsible for the statistical deviance.

An item is said to be biased when groups with the same ability do not have the same probability of correctly responding to the item. In our study, the Mantel-Haenszel statistic (MH; Holland & Thayer, 1988) was chosen for the detection of biased items. In a review of statistical approaches for assessing measurement bias (Millsap & Everson, 1993), the MH method was cited as one of the more widely used methods for detecting item-level measurement bias.

Even if total scores on subtests might strongly reflect the same dimensions in different groups, individual items can still be biased. Literature on cancellation of differential item functioning (DIF) indicated that biased items may partially compensate for each other (Drasgow, 1987; Nandakumar, 1993). On the same account, comparability of dimensions is no guarantee for the absence of quantitative item bias.

The analyses were conducted on the GATB subtests Computation, Three-Dimensional Space, and Vocabulary. We had to choose which items would be analyzed with the MH method. If all of the items in the subtest are analyzed, the items at the end of the subtest will be statistically biased, due to differences in the number of items completed by each group. However, this statistical bias can be understood as *position bias.* Sound conclusions can only be drawn for the items that have been completed by a sufficient number of participants. For this reason, only items completed by close to 90% of the immigrants were included in the analyses. The tests can then be considered to resemble power tests.

Because of the relatively large size of the groups in this research, small differences in p values could have easily resulted in significant χ^2 values. Tests for statistical significance were therefore conducted with an alpha level of .01. We decided to test for small differences between groups: The mean difference in p values of an item for the two groups for each of the scoring categories had to be larger than .05 (this corresponds with a difference between the two groups of about .01 *SD*). When the MH method is used, one must decide how many items are to be analyzed. A rule of thumb is that a minimum of 10 items is needed to calculate the total score for the nonbiased items. The MH method takes the total score as an estimate for the position on the latent trait. If this total score is calculated on the basis of fewer than 10 items, the reliability of the estimate will be too low. For that reason, Arithmetic Reason was not analyzed because 90% of the immigrants finished fewer than 10 items. As a final step, the direction of the bias was taken into account. Was the item statistically biased against immigrants or against majority group members, or were the differences in means within the different score categories in opposite directions, so that the item could not be classified as being biased against a specific group?

After we identified the biased items, our next task was to explain the statistical deviance. We conducted post hoc inspection of the statistically biased items, with the emphasis on identifying striking characteristics that could be related to well-known differences between majority and immigrant groups, such as differences in cultural background, status in society, and Dutch language proficiency. Finally, we estimated the effect of the biased items on the mean scores of the immigrants.

Spearman's hypothesis. The correlation between the g loading and the difference between the means for the immigrant and majority groups was calculated for each immigrant group so that the proposition that the differences between the groups are attributable to a general difference in capacities could be tested. It has been shown that the correlation between the g loading, computed by means of hierarchical factor analysis and the difference between the means for Whites and Blacks in the United States is an empirical fact (Jensen, 1985). For the aptitudes measured by the GATB for groups of Whites and Blacks, a

Pearson r of .71 has been reported (Jensen, 1985). This means that General Intelligence or g is the predominant factor, but not the sole factor, determining the size of the differences between the two groups. Jensen (1993) states that seven methodological requirements for the testing of what he called *Spearman's hypothesis* have to be met:

1. The samples should not be selected on any highly g-loaded criteria.
2. The variables should have reliable variation in their g loadings.
3. The variables should measure the same latent traits in all groups. The congruence coefficient of the factor structure should have a value of >.85.
4. The variables should measure the same g in the different groups; the congruence coefficient of the g values should be >.95.
5. The g loadings of the variables should be determined separately in each group. If the congruence indicates a high degree of similarity, the g loadings of the different groups should be averaged.
6. To rule out the possibility that the correlation between the vector of g loadings (V_g) and the vector of mean differences between the groups, or effect sizes (V_{ES}), is strongly influenced by the variables' differing reliability coefficients, V_g and V_{ES} should be corrected for attenuation by dividing each value by the square root of its reliability.
7. The test of Spearman's hypothesis is the Pearson correlation (r) between V_g and V_{ES}. To test the statistical significance of r, Spearman's rank order correlation (r_s) should be computed and tested for significance.

The g loadings were computed, using the first unrotated factor of a principal axis factor analysis (Jensen & Weng, 1994). Because of the limited sampling of broad abilities of the GATB, it is not optimal for a precise and theoretically sound estimate of g loadings. As an example, although Carroll (1993) shows that Visualization tests generally have high g loadings and tests of General Psychomotor Speed generally have low g loadings, Table 2 shows that the subtests Mark Making and Three-Dimensional Space differ little in their g loadings in the majority group. The empirical g loadings will therefore only be used to check the comparability of g loadings for majority groups and immigrants.

The best estimate of the g loadings was found in a factor analytic study of the Dutch version of the GATB 1002 A with a large number of other tests, using the first unrotated factor of a principal axis factor analysis (Dutch GATB Manual; van der Flier & Boomsma-Suerink, 1994, p. 51). Table 2 shows that these g loadings are similar to the g loadings reported by Carroll (1993): the highest values for Crystallized and Fluid tests, somewhat lower values for tests that measure Broad Visual Perception, and a low value for Psychomotor tests. These estimated values of the g loadings were used for the correlation of V_g and V_{ES} for the four comparisons. This procedure departs somewhat from Jensen's fifth requirement, but in this case it seems preferable.

To check whether the correlation between V_g and V_{ES} was due to g or to other factors, such as bias, the regression of the standardized mean group differences (D) on the estimated g loadings was computed and is shown later in graphical form for the Turks. If a test shows a greater difference between two groups than can be expected on the basis of its g loadings, this may mean that it is biased, for instance, that the test is measuring specific knowledge or a different ability construct.

Table 2
Empirical g Loadings (g) and Estimated g Loadings (g_{est}) for Majority Group

Subtest	g	g_{est}
Vocabulary	.69	.68
Arithmetic Reason	.74	.68
Computation	.69	.67
Name Comparison	.76	.62
Three-Dimensional Space	.46	.58
Tool Matching	.55	.49
Form Matching	.58	.53
Mark Making	.40	.14

Note. Estimated g loadings are from the Dutch General Aptitude Test Battery Manual (van der Flier & Boomsma-Suerink, 1994).

Results

Means

Table 3 shows the mean scores of the majority group and data from the Dutch GATB Manual (van der Flier & Boomsma-Suerink, 1994, p. 149). This manual gives the mean scores for a representative sample of pupils in the whole range of secondary education for non-learning-disabled students in the Netherlands. The pupils were all in their last year of secondary education, the majority between 16 and 18 years of age, depending on the type of school. The data were collected before 1982. The standard deviations in the two groups are strongly comparable, and the means in the majority group are, on average, about one third of a standard deviation lower.

Table 4 shows that the scores of the immigrants on the subtests were, on the average, about one standard deviation lower than the scores of the majority group. On the whole, North Africans and Turks scored lower than Surinamese and Antillians, especially on tests with a verbal component. The subtests that call for knowledge of the Dutch language showed the largest mean differences between the majority group and each immigrant group. This was most evident for the subtest Vocabulary and to a lesser extent for the subtests Arithmetic Reasoning and Name Comparison. Tests that call less strongly for knowledge of the Dutch language, such as Computation, Tool Matching, Three-Dimensional Space, and Form Matching, however, also showed considerable differences between the majority group and the immigrant groups. The smallest difference was found for the subtest Mark Making, which measures General Psychomotor Speed.

Table 3
Means and Standard Deviations of the General Aptitude Test Battery (GATB) Subtests for the Majority, Norm, and Test–Retest Groups

	Majority		Norm		Test–retest
Subtest	*M*	*SD*	*M*	*SD*	*SD*
Vocabulary	24.23	6.69	25:94	7.26	4.61
Arithmetic Reason	13.06	3.39	14.19	3.35	2.91
Computation	21.02	4.81	24.68	4.57	3.86
Name Comparison	60.04	12.60	67.69	15.74	11.95
Three-Dimensional Space	20.88	5.67	21.55	5.50	4.74
Tool Matching	28.69	5.57	32.35	5.97	5.81
Form Matching	28.40	6.03	28.76	5.90	4.60
Mark Making	68.46	9.98	72.93	9.88	9.40

Note. For a description of the norm group, see text. Test–retest group from Dutch GATB Manual; *SD*s for the test–retest group are from Bosch (1973).

Reliabilities

The percentage of Surinamese that answered specific items at the beginning of the subtests was highly comparable to the percentage of Antillians that answered the same specific items. Therefore, the same number of items was chosen for the two groups to constitute a power test, on which the alpha was computed. On the same grounds, the same was done for the North Africans and Turks. Alphas were therefore computed on the basis of items that were answered by close to 90% of the participants in a group.

Table 5 shows that reliability coefficients were higher for the immigrant groups in practically all cases. Tables 3 and 4 show that, in most cases, the standard deviations were higher for the immigrant than for the majority group, so the true variance is probably larger in the immigrant groups.

However, although the analysis was limited to those items that were answered by close to 90% of the immigrants, the problem remained that the immigrants answered fewer items than did the majority group at the end of the test. Therefore, there were more immigrants with zero scores on relatively more items, which influenced the correlations and somewhat inflated the alphas. It would therefore appear that the higher values of coefficient alpha for the immigrants could partly be explained as an artifact.

Dimensional Comparability

The comparability of the dimensions of the subtests for the majority groups and the immigrant groups was investigated by means of structural equation modeling (EQS); subtest correlations are presented in Tables 6, 7, and 8. The covariance matrix of the majority group was compared, in separate analyses, with the covariance matrices of the Surinamese, the Antillians, the North Africans, and the Turks. Table 9 shows that the values of the CFI varied between .976 and .994. From this it may be concluded that the two covariance matrices in the four different comparisons are essentially identical.

The fit was further explored by fitting increasingly constrained models to the data. In the first analysis, a test of

Table 4
Deviations From the Mean of the Majority Group Expressed in Standard Deviations of the Majority Group (Dev.) and Standard Deviations (SD) on the General Aptitude Test Battery Subtests by Immigrant Group

	Surinamese		Antillians		North Africans		Turks	
Subtest	Dev.	*SD*	Dev.	*SD*	Dev.	*SD*	Dev.	*SD*
Vocabulary	1.05	6.84	1.32	6.97	2.07	5.00	1.96	5.60
Arithmetic Reason	1.01	3.43	1.05	3.03	1.76	3.14	1.24	3.37
Computation	0.40	5.11	0.52	5.00	1.00	5.41	0.80	5.13
Name Comparison	0.64	13.29	0.74	15.64	1.40	11.39	0.96	12.37
Three-Dimensional Space	1.06	5.62	0.91	6.42	1.43	5.59	1.03	6.00
Tool Matching	0.83	6.17	0.66	7.23	0.90	6.45	0.57	6.23
Form Matching	0.65	6.27	0.61	6.94	0.87	6.53	0.53	6.20
Mark Making	0.01	11.71	0.14	13.35	0.18	12.47	0.29	11.98

Table 5
Values of Reliability Coefficients (Alphas) for Majority Group Members and Immigrants

Subtest	Group			Group		
	Majority	Surinamese	Antillians	Majority	North Africans	Turks
Three-Dimensional Space	.76	.80	.79	.78	.82	.82
Vocabulary	.69	.76	.76	.64	.60	.64
Arithmetic Reason	.46	.59	.62	.36	.63	.63
Computation	.64	.71	.73	.47	.70	.68

Note. Alphas were computed on the basis of items that were answered by close to 90% of the participants in a group. For the Surinamese and the Antillians, 15 out of 40 items of Three-Dimensional Space, 16 out of 50 items for Vocabulary, 8 out of 25 items for Arithmetic Reason, and 16 out of 50 items for Computation were analyzed. For the North Africans and the Turks, 16 out of 40 items for Three-Dimensional Space, 12 out of 50 items for Vocabulary, 7 out of 25 items for Arithmetic Reason, and 12 out of 50 items for Computation were analyzed.

the same number of factors was examined. In addition to zero loadings, three loadings with a value of one were fixed in order to circumvent problems of identifiability, namely, the loading of Computation on G_H, of Tool Comparison on G_V, and of Mark Making on G_P. Because a hierarchical model was tested, the factors were oblique. Table 9 shows that the equal factor model gave a good fit to the data. Additional analyses revealed that the data showed a good fit for the majority group, $\chi^2(17, N = 800) = 233.88, p < .001$, CFI = .908; the Surinamese, $\chi^2(17, N = 523) = 130.86, p < .001$, CFI = .931; the North Africans, $\chi^2(17, N = 155) = 55.64, p < .001$, CFI = .905; the Turks, $\chi^2(17, N = 262) = 96.44, p < .001$, CFI = .917; and fairly well for the Antillians, $\chi^2(17, N = 122) = 65.18, p < .001$, CFI = .853. The loadings of the factor solution are reported in Table 10. The loading of Mark Making on the factor General Psychomotor Speed is not reported because the value is 1 in every group.

Table 9 shows that the fit decreased for the Surinamese, Antillians, and Turks when the factor loadings were held equal across the groups; a significant increase of chi-square value took place for all of these groups: for the Surinamese, $\Delta\chi^2(6, N = 1{,}323) = 19.18, p = .0039$; the Antillians, $\Delta\chi^2(6, N = 922) = 19.62, p = .0032$; the North Africans, $\Delta\chi^2(6, N = 955) = 2.56, p = .86$; and the Turks, $\Delta\chi^2(6, N = 1{,}062) = 53.25, p = .00001$. Although there are significant effects, the fit indices are so high that the differences are probably very small. In sum, the fit of the model postulating the same numbers of factors for the different groups was adequate and a model of equal factor loadings was also adequate, albeit slightly less so. The dimensions of the subtests for the majority group and the immigrant groups were highly comparable.

The values of the congruence coefficient were high, varying from .970 to .997. Values above .95 are generally considered to be an indication of strong similarity between the various factor structures. For each factor, the values of the congruence coefficient were higher than the values between .83 and .92 reported by Vandenberg and Hakstian (1978).

Table 10 shows the factor loadings; the discrepancies in these loadings indicate that some of the subtests measure something different in the immigrant group than in the majority group or, in other words, that the subtests differ in their construct validity. The factor solutions of the

Table 6
Correlations Between General Aptitude Test Battery Subtest Scores for the Majority Group

Subtest	1	2	3	4	5	6	7	8
1. Vocabulary	—							
2. Arithmetic Reason	.564	—						
3. Computation	.475	.740	—					
4. Name Comparison	.558	.556	.530	—				
5. Three-Dimensional Space	.346	.297	.211	.210	—			
6. Tool Matching	.311	.258	.254	.439	.422	—		
7. Form Matching	.322	.294	.280	.415	.512	.529	—	
8. Mark Making	.259	.223	.267	.391	.080*	.252	.266	—

Note. $N = 806$.
* $p < .05$; all other correlations, $p < .01$.

Table 7
Correlations Between General Aptitude Test Battery Subtest Scores for the North Africans and the Turks

Subtest	1	2	3	4	5	6	7	8
1. Vocabulary	—	.652	.464	.574	.493	.379	.359	.409
2. Arithmetic Reason	.553	—	.729	.603	.508	.450	.426	.305
3. Computation	.354	.592	—	.566	.417	.471	.412	.294
4. Name Comparison	.569	.432	.380	—	.422	.609	.442	.447
5. Three-Dimensional Space	.328	.248	.184	.325	—	.512	.549	.187
6. Tool Matching	.357	.280	.224	.557	.524	—	.573	.304
7. Form Matching	.306	.238	.314	.480	.554	.535	—	.233
8. Mark Making	.328	.233	.248	.430	.107[a]	.195	.281	—

Note. n = 167 for the North Africans; n = 275 for the Turks. Values for the North Africans are below the diagonal; values for the Turks are above the diagonal.
[a] Nonsignificant correlation; all other correlations, $p < .01$.

Antillians and the North Africans showed substantial discrepancies (changes of >.10) for the subtests Arithmetic Reason, Computation, and Name Comparison, and for the Antillians for Tool Matching as well. The Surinamese showed substantial discrepancies for the subtests Name Comparison and Tool Matching. The Turks showed no substantial discrepancies. Therefore the construct validity of these subtests differs for most of the immigrant groups but, taken as a whole, the differences of the factor structures between the immigrant groups and the majority groups are small. The only subtest that shows discrepancies in factor loadings in three of the four immigrant groups is Name Comparison. No other clear pattern seems detectable in the discrepancies: Some subtests with a language component do not change their loadings, whereas some subtests without language components do change; Surinamese and Antillians show more differences with the majority group than do Turks.

Differential Item Functioning

Table 11 displays the items identified as being statistically biased (showing DIF) against the immigrant groups, using the Mantel-Haenszel statistic. Post hoc inspection of statistically biased items in Vocabulary was conducted, with emphasis on the identification of striking characteristics. In the statistically biased items, words were found that could be interpreted as being difficult or old-fashioned for immigrants. It seems that the subtest calls more strongly on knowledge of the Dutch language than might be desired, given what the test is supposed to measure. Stated in other words, the test seems to call more strongly on Lexical Knowledge and less strongly on Induction.

Other items containing words that were seemingly of a comparable degree of difficulty turned out not to be statistically biased. This is in line with the conclusion drawn in previous item-bias research that item bias is not always predictable. In our research, the appearance of statistical bias at the end of the test might, in some cases, be an artifact of the method used, in which case the statistical bias can be interpreted as position bias.

Table 11 also displays the effect of statistically biased items on the scores of the immigrants. On the subtest Vocabulary, for North Africans, the mean difference between the p values of the five statistically biased items

Table 8
Correlations Between General Aptitude Test Battery Subtest Scores for the Surinamese and the Antillians

Subtest	1	2	3	4	5	6	7	8
1. Vocabulary	—	.434	.371	.486	.493	.399	.369	.349
2. Arithmetic Reason	.530	—	.580	.261	.403	.267	.227	.172*
3. Computation	.436	.723	—	.376	.240	.376	.356	.248
4. Name Comparison	.498	.521	.528	—	.254	.625	.455	.490
5. Three-Dimensional Space	.268	.252	.232	.250	—	.437	.467	.174*
6. Tool Matching	.342	.358	.365	.595	.464	—	.638	.374
7. Form Matching	.325	.312	.341	.483	.523	.590	—	.321
8. Mark Making	.318	.312	.358	.485	.135	.364	.308	—

Note. n = 535 for the Surinamese; n = 126 for the Antillians. Values for the Surinamese are below the diagonal; values for the Antillians are above the diagonal.
* $p < .025$; all other correlations, $p < .01$.

Table 9
Results of the Structural Equation Modeling Comparing the Majority Group With Each Immigrant Group

Model	χ^2	*df*	$\Delta\chi^2$	CFI
Majority and Surinamese				
Equal covariance matrices	58.07*	36		.994
Equal factor models	364.74	34		.918
Equal factor loadings	283.92	40	19.18**	.914
Majority and Antillians				
Equal covariance matrices	100.35	36		.976
Equal factor models	299.06	34		.902
Equal factor loadings	318.68	40	19.62**	.897
Majority and North Africans				
Equal covariance matrices	98.80	36		.977
Equal factor models	289.52	34		.908
Equal factor loadings	292.08	40	2.56[a]	.909
Majority and Turks				
Equal covariance matrices	99.69	36		.981
Equal factor models	288.14	34		.924
Equal factor loadings	341.39	40	53.25	.909

Note. CFI = comparative fit index.
[a] Not significant.
* $p < .05$. ** $p < .01$. All others, $p < .001$.

for the two groups for each of the four scoring categories varied between .06 and .16. This means that if the five statistically biased items were replaced by five nonbiased items, then the North Africans could be expected to have, on average, 0.532 of an item more correct, which would yield a score about 0.11 *SD* higher. Replacement of the statistically biased items for the Surinamese would yield a score about 0.04 *SD* higher. It therefore appears that the statistically biased items led to somewhat depressed scores. Because not all items were analyzed with the MH statistic, potential bias in the remaining, nonanalyzed items could have gone undetected; these values of the effect of statistically biased items on the scores of immigrants are therefore likely to be underestimates.

Because only a limited set of items was analyzed, measurement error in the ability estimates could have an effect on the item-bias analyses. However, corrections for this potential error are generally not carried out because the effects are usually expected to be small. In this study, because of the relatively high reliabilities, the measurement error in the ability estimates will probably not have a strong effect either.

Spearman's Hypothesis

Jensen's methodological requirements were met. The samples in this study varied from train cleaner to rail maintenance expert and were not selected on any highly *g*-loaded criteria, so there is no indication that the *g* variance in the samples is markedly restricted. Table 12 shows that the subtests have reliable variation in their estimated *g* loadings, with a range from .14 to .68. The third requirement is that the tests should measure the same latent traits in the various groups; it was clear from the preceding analyses that this was the case. A comparison of the empirical *g* loadings of the majority group with the empirical *g* loadings of the four immigrant groups resulted in values of the congruence coefficient that varied between .978 and .995. The empirical *g* loadings are therefore highly comparable in the different groups.

We used the estimated *g* loadings for the correlation of V_g and V_{ES} (r is Pearson correlation; r_s = Spearman's rank order correlation): for the Surinamese, $r = .72$, $r_s = .42$, $p = .151$, and $D = 1.47g - 0.10$; for the Antillians, $r = .77$, $r_s = .71$, $p = .025$, and $D = 1.54g - 0.10$; for the North Africans, $r = .84$, $r_s = .87$, $p = .003$, and $D = 2.77g - 0.32$; and for the Turks, $r = .70$, $r_s = .87$, $p < .003$, and $D = 2.02g - 0.18$.

To check the influence of differing reliability coefficients, it is best to use test–retest reliabilities. Table 12 shows the values from the Dutch GATB Manual (van der Flier & Boomsma-Suerink, 1994, p. 114). Table 3 shows that the variability in the sample in the manual was lower than the reliability in the current sample. To adjust for

Table 10
Factor Loadings for Majority and Immigrant Groups, and Congruence Coefficients for Comparison With Majority Group

	Majority		Surinamese		Antillians		North Africans		Turks	
Subtests	G_H	G_V	G_H	G_V	G_H	G_V	G_H	G_V	G_H	G_V
Vocabulary	.65		.61		.65		.73		.72	
Arithmetic Reason	.87		.86		.69		.75		.90	
Computation	.82		.82		.71		.62		.78	
Name Comparison	.54	.28	.41	.56	.19	.57	.46	.40	.45	.37
Three-Dimensional Space		.60		.56		.53		.67		.69
Tool Matching		.70		.83		.86		.77		.79
Form Matching		.79		.74		.73		.75		.72
Congruence coefficient			⟨.997⟩	⟨.975⟩	⟨.972⟩	⟨.970⟩	⟨.989⟩	⟨.994⟩	⟨.997⟩	⟨.992⟩

Note. G_H = Hybrid of Fluid and Crystallized Intelligence; G_V = Broad Visual Perception.

Table 11
Differential Item Functioning Using the Mantel-Haenszel Statistic

	Surinamese			Antillians			North Africans			Turks		
Subtest	No.	Bias	Effect	No.	Bias	Effect	No.	Bias	Effect	No.	Bias	Effect
Vocabulary	16	6, 7	0.04	16	2, 6, 7, 10, 12	0.07	12	2, 6, 8, 10, 12	0.11	12	2, 8, 10	0.06
Three-Dimensional Space	15	3, 10, 14	0.04	15	10, 14	0.02	16	2, 10, 12, 14	0.07	16	12, 14	0.05
Computation	16	14	0.01	16	—	—	12	11	0.01	12	10	0.01

Note. The Mantel-Haenszel statistic is computed on the basis of items answered by close to 90% of the participants in a group. No. = number of items analyzed; Bias = biased items; Effect = estimated score improvement (effect) if the biased items were replaced with nonbiased items.

these differences, Gulliksen's (1950, p. 124) adjustment formula was used; adjusted reliabilities are reported in Table 12. Table 5 shows the alphas for the majority group for four tests; when computed on the items that were finished by approximately 90% of the research participants from the majority group, the values are reasonably comparable to the corresponding values of r_{xx}. The alphas of the majority group and the immigrants do not differ much in most cases, so the same reliability coefficients are used in all of the groups. Each value in V_g and V_{ES} was corrected for attenuation, and the correlation between the disattenuated vectors was computed: $r = .76$ for the Surinamese, $r = .78$ for the Antillians, $r = .82$ for the North Africans, and $r = .64$ for the Turks. The results of these analyses demonstrate that the substantiation of Spearman's hypothesis is not an artifact of variation in reliability of the GATB subtests.

Thus, *g* is the predominant factor, but not the sole factor, determining the size of the differences between the majority group and the immigrant groups.

Although the regression of *D* on *g* gives a good fit to the eight data points for the Surinamese, some subtests show a larger *D* than would be expected on the basis of their *g* loadings, whereas other subtests show a smaller *D* than expected. Subtests that are above the regression line may be relatively more difficult for immigrants, and subtests that are below the regression line may be relatively easier. The two subtests that are above the regression line are Vocabulary (+.15 *SD*) and Arithmetic Reason (+.11 *SD*) and both have a verbal component. An interpretation of this high value of *D* in terms of language bias is supported by a small biasing effect at the level of the items in the subtest Vocabulary. We have already stated that this biasing effect is most likely to be an underestimate. No support is found in the EQS analyses because Vocabulary and Arithmetic Reason show no changes in construct validity. On the other hand, the subtests Three-Dimensional Space and Tool Matching are above the regression line, whereas their items have no verbal component.

Figure 1 shows that the regression of *D* on *g* gives a good fit to the eight data points of the Turks; the regression

Table 12
Differences in Means Between Majority and Immigrant Groups in Sigma Units (D), Empirical g Loadings (g), Estimated g Loadings for Majority Group (g_{est}), Test–Retest Reliability, Adjusted Test–Retest Reliability and Congruence Coefficients Between g Loadings of Majority and Immigrants

	Majority		Surinamese		Antillian		North Africans		Turks			
Subtest	*g*	g_{est}	*D*	*g*	*D*	*g*	*D*	*g*	*D*	*g*	r_{xx}	r'_{xx}
Vocabulary	.69	.68	1.05	.61	1.32	.67	2.07	.67	1.96	.70	.65	.83
Arithmetic Reason	.74	.68	1.01	.69	1.05	.52	1.76	.61	1.24	.80	.69	.77
Computation	.69	.67	.40	.69	.52	.57	1.00	.55	.80	.72	.71	.81
Name Comparison	.76	.62	.64	.79	.74	.70	1.40	.78	.96	.79	.83	.85
Three-Dimensional Space	.46	.58	1.06	.46	.91	.57	1.43	.54	1.03	.65	.75	.83
Tool Matching	.55	.49	.83	.69	.66	.75	.90	.65	.57	.70	.52	.48
Form Matching	.58	.53	.65	.63	.61	.67	.87	.63	.53	.62	.52	.72
Mark Making	.40	.14	.01	.51	.14	.49	.18	.42	.29	.44	.83	.85
Congruence coefficient		.984		.994		.978		.991		.995		

Note. Estimated *g* loadings and test–retest reliability are from the Dutch General Aptitude Test Battery Manual (van der Flier & Boomsma-Suerink, 1994).

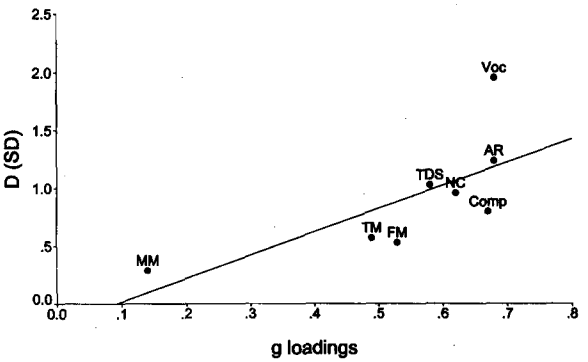

Figure 1. Regression of the standardized mean difference (D) between the majority group and the Turks in *SD* units on the estimated *g* loadings. Voc = Vocabulary; AR = Arithmetic Reason; Comp = Computation; NC = Name Comparison; TDS = Three-Dimensional Space; TM = Tool Matching; FM = Form Matching; MM = Mark Making.

equation is $D = 2.02g - 0.18$. Vocabulary is .77 *SD* above the regression line, whereas Arithmetic Reason and Three-Dimensional Space are only slightly above the regression line. An interpretation in terms of language bias is supported by the item-bias analyses, but not by the EQS analyses. It might be, however, that Vocabulary changes its construct validity in the Turkish group but that this change cannot be detected with these data. The low mean score of Turks on Vocabulary might be decomposed into the following: (a) a relatively low mean level of *g*, (b) a relatively low mean level of G_f or G_c (or both), (c) a relatively low mean level of proficiency in Dutch, and (d) bias that is not explained by (c).

Discussion

The results provide important indications that the test scores of immigrants and majority group members reflect the same dimensions. The tests measure, to a large extent, the same dimensions at the level of the total score, with the exception of the subtest Name Comparison that shows a strong group difference of the construct validity. The GATB subtests are also largely comparable at the level of individual items.

Means and Reliabilities

The subtests that call for knowledge of the Dutch language showed the largest differences between the means of the majority group and the immigrants. To a lesser extent, however, tests that do not call for knowledge of the Dutch language also showed a large difference between the scores of the majority group and the immigrants. The assumption that the differences on the subtests can only be attributed to differences in proficiency in Dutch is not, therefore, in accordance with these outcomes.

The measures of reliability have a higher value for the immigrant groups, but this can be explained by higher test score variance.

Dimensional Comparability

Comparisons of the covariance matrices by means of structural models (EQS) showed that they were essentially identical. The data gave a good fit to Carroll's version of the Cattell-Gustafsson model for most of the groups, and the factor solution of the majority group was also quite strongly evident in the immigrant groups. The values of the congruence coefficient were high, and for every factor they were higher than the values found in cross-cultural research. This indicates a strong resemblance between the different factor structures for the groups included in the present research. The strong similarity of *g* loadings provides additional evidence for strong dimensional comparability.

The factor solutions of the different groups showed discrepancies in the loadings of some subtests. This points to a difference in construct validity. The discrepancies in factor loadings of the subtest Name Comparison for three of the immigrant groups suggests that some of the immigrants had problems with the names used in the subtests; these names are, notwithstanding an international tinge, for the most part European. In the majority group this subtest distinguishes, among other things, the speed and accuracy of the formation of the word image. Given the substantial loadings seen on the Broad Visual Perception factor, it looks as though immigrants fell back on the comparison of collections of loose signs.

Although some of the other loadings differ from the loadings in the majority group, there is no readily recognizable pattern to be found. The subtest Vocabulary shows the same loading as in the majority group; although it might be that the subtest has somewhat less the character of a test for Fluid Intelligence and somewhat more the character of a test for Crystallized Intelligence, our data cannot answer this question. Note, however, that the *g* loading of the test remains high in every immigrant group.

In short, although there were discrepancies in the loadings of some subtests, the dimensional structures were largely the same.

Differential Item Functioning

The subtests Vocabulary and Three Dimensional Space contain many statistically biased items, whereas Computation contains only one at most. Inspection of the statistically biased items in Vocabulary and Arithmetic Reason reveals that they mostly contain relatively difficult words.

It seems justifiable to conclude that, in most cases, bias arises in those items that contain words which fall outside the vocabulary of some of the immigrants. The item bias in Vocabulary appears to have lowered the mean score of the immigrants by at least one tenth of a standard deviation. It appears that two out of three of the analyzed subtests are not comparable at the level of individual items. Item bias had no large influence on the mean scores.

Spearman's Hypothesis

Of the different explanations for the differences in means between the groups, Spearman's hypothesis received the strongest support. For all four of the immigrant groups, *g* is the predominant factor accounting for differences between the majority group and the immigrant groups. The non-perfect *r* may be explained by (a) different intelligence profiles, (b) sampling error in *g* loadings, (c) biasing factors, and (d) measurement error.

Conclusions

In sum, the group differences with respect to the construct validity at the level of total scores and at the level of individual items were not large. Nevertheless, these small group differences do lower the construct validity of the GATB subtests and blur their representation of the capacities of the immigrants. These small group differences could have large consequences.

Because there are no data at our disposal with which to assess predictive validity in the Dutch situation, the possibility that some of the differences in means were caused by non-test-specific abilities cannot be ruled out. We cannot say with certainty to what extent the large differences in scores were caused by a lower level of aptitudes in the immigrant group and to what extent were caused by bias in the test.

The finding in the American literature of no test bias against minorities is not perfectly, but strongly confirmed in this study. A review of studies in the Netherlands on test bias (te Nijenhuis, 1997) strongly supports the findings from this study.

A practical conclusion is that the test can be put to good use for comparisons within culturally homogeneous groups of non-native-born, non-native-language minorities. Use of the test for comparisons between immigrants and majority group members, however, requires supplementary research.

References

Arvey, R. D., & Faley, R. H. (1988). *Fairness in selecting employees* (2nd ed.). Reading, MA: Addison-Wesley.

Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles, CA: BMDP Statistical Software.

Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux [New methods for diagnosing the intellectual level of abnormals]. *Année Psychologique, 11,* 191–336.

Bosch, F. (1973). *Inventarisatie, beschrijving en onderzoek m.b.t. de wijzigingen (1971) van de G.A.T.B., incl. test–hertest onderzoek* [Survey, description, and research with regard to the changes (1971) of the GATB, including test–retest research.]. Utrecht, the Netherlands: Dutch Railways.

Burt, C. (1948). The factorial study of temperamental traits. *British Journal of Psychology, 1,* 178–203.

Carretta, T. R., & Ree, M. J. (1995). Near identity of cognitive structure in sex and ethnic groups. *Personality and Individual Differences, 19,* 149–155.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge University Press.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action.* Amsterdam: North Holland.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Evers, A., van Vliet-Mulder, J. C., & Ter Laak, J. (1992). *Documentatie van tests en testresearch in Nederland* [Documentation of tests and test research in the Netherlands]. Assen, the Netherlands: Van Gorcum.

Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist, 49,* 955–963.

Gulliksen, H. (1950). *Theory of Mental Tests.* New York: Wiley.

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8,* 179–203.

Gustafsson, J.-E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 35–71). Hillsdale, NJ: Erlbaum.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Herrnstein, R. J., & Murray, C. (1994). *The Bell curve: Intelligence and class structure in American life.* New York: Free Press.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Hunter, J. E. (1983a). *The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance for the U.S. Employment Service* (USES Test Research Rep. No. 44). Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Hunter, J. E. (1983b). *Fairness of the General Aptitude Test Battery: Ability differences and their impact on minority hiring rates* (USES Test Research Rep. No. 46). Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential

validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 87,* 721–735.

Jensen, A. R. (1980). *Bias in mental testing.* London: Methuen.

Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 193–219.

Jensen, A. R. (1993). Spearman's hypothesis tested with chronometric information-processing tasks. *Intelligence, 17,* 47–77.

Jensen, A. R., & Weng, L.-J. (1994). What is a good *g*? *Intelligence, 18,* 231–258.

Jöreskog, K. G. (1969). A general approach to the confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education, MacMillan.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30,* 293–311.

te Nijenhuis, J. (1997). *Comparability of test scores for immigrants and majority group members in the Netherlands.* Doctoral dissertation, Vrije Universiteit, Amsterdam. Manuscript in preparation.

Tucker, L. R. (1951). A method for synthesis of factor analysis studies (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

U.S. Department of Labor. *Manual for the General Aptitude Test Battery* (1970). Washington, DC: Author.

van der Flier, H., & Boomsma-Suerink, J. L. (1994). *Handboek GATB* [GATB manual]. Amsterdam: Stichting G.A.T.B.-Research.

Vandenberg, S. G., & Hakstian, A. R. (1978). Cultural influences on cognition: A reanalysis of Vernon's data. *International Journal of Psychology, 13,* 251–279.

Received June 24, 1996
Revision received May 20, 1997
Accepted May 23, 1997 ■